NPSCS-93-007

NAVAL POSTGRADUATE SCHOOL
Monterey, California

(2)

Inferring Depictions in Natural-Language Captions for Efficient Access to Picture Data

Neil C. Rowe[1]

July 1993

TECHNICAL REPORT

October 1, 1992 to July 1993

Approved for public release; distribution is unlimited.

Prepared for:

Naval Postgraduate School
Monterey, California 93943

NAVAL POSTGRADUATE SCHOOL

Monterey, California

REAR ADMIRAL T. A. MERCER
Superintendent

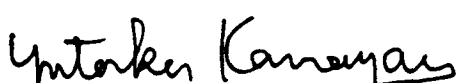
HARRISON SHULL
Provost

This report was prepared with research funded by the Naval Research Funds provided by the Naval Postgraduate School.

This report was prepared by:

NEIL C. ROWE
Associate Professor of
Computer Science

Reviewed by:



YUTAKA KANAYAMA
Associate Chairman for
Research

Released by:

PAUL MARTO
Dean of Research

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NPSCS-93-007	5. MONITORING ORGANIZATION REPORT NUMBER(S) Naval Postgraduate School
6a. NAME OF PERFORMING ORGANIZATION Computer Science Dept. Naval Postgraduate School 6b. OFFICE SYMBOL (if applicable) CS	7a. NAME OF MONITORING ORGANIZATION ONR
6c. ADDRESS (City, State, and ZIP Code) Monterey, CA 93943	7b. ADDRESS (City, State, and ZIP Code) San Diego, CA
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Naval Postgraduate School 8b. OFFICE SYMBOL (if applicable) NPS	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER DARPA 13 Project under AO 8939
8c. ADDRESS (City, State, and ZIP Code) Monterey, CA 93943	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Inferring depictions in natural-language captions for efficient access to picture data

12. PERSONAL AUTHOR(S)
Neil C. Rowe

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM 9210 TO 9307	930719	23

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	caption, multimedia, information retrieval, natural language, parsing, reference, Prolog

19. ABSTRACT (Continue on reverse if necessary and identify by block number)
Multimedia data can require significant examination time to find desired features ("content analysis"). An alternative is using natural-language captions to describe the data, and matching captions to English queries. But it is hard to include everything in the caption of a complicated datum, so significant content analysis may still seem required. We discuss linguistic clues in captions, both syntactic and semantic, than can simplify or eliminate content analysis. We introduce the notion of concept depiction and rules for depiction inference. Our approach is implemented in an expert system which demonstrated significant increases in recall in experiments.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL Neil C. Rowe	22b. TELEPHONE (Include Area Code) (408) 656-2462 22c. OFFICE SYMBOL CSRp

Inferring depictions in natural-language captions for efficient access to picture data

Neil C. Rowe[1]

Department of Computer Science

Code CS/Rp, U. S. Naval Postgraduate School

Monterey, CA USA 93943

(rowe@cs.nps.navy.mil)

ABSTRACT

Multimedia data can require significant examination time to find desired features ("content analysis"). An alternative is using natural-language captions to describe the data, and matching captions to English queries. But it is hard to include everything in the caption of a complicated datum, so significant content analysis may still seem required. We discuss linguistic clues in captions, both syntactic and semantic, that can simplify or eliminate content analysis. We introduce the notion of concept depiction and rules for depiction inference. Our approach is implemented in an expert system which demonstrated significant increases in recall in experiments.

[1] This work was sponsored by the Naval Ocean Systems Center in San Diego, California, the Naval Air Warfare Center, Weapons Division, in China Lake, California, the U. S. Naval Postgraduate School under funds provided by the Chief for Naval Operations, and the Defense Advanced Research Projects Administration.

Keywords: information retrieval, multimedia, captions, databases, natural language, focus, denotation, parsing, cooperativeness, man-machine interfaces

1. Introduction

Multimedia data such as pictures, audio, and video pose a fundamental challenge for information retrieval because they can contain far more information than can be indexed feasibly, yet content searching can be so difficult. Natural-language captions can summarize complex multimedia data, and these captions can be examined first before retrieving the (often-bulky) data. But when should we do a content analysis of data if a user asks about things not specified in a caption?

This problem has arisen in our development of a multimedia database system that uses English captions in an integral way to provide quick access to relevant pictures, as summarized in Rowe and Guglielmo (1993). Our captions are not just incidental information for a media datum, but the primary indexing method for data. This requires parsing and semantic interpretation of natural language, resulting in a "meaning list" representing each caption as a list of logical records. For example, the caption "pylon of an F-18 wing" can be represented by the meaning list "a_kind_of(X,pylon) and belongs_to(X,Y) and a_kind_of(Y,wing) and part_of(Y,Z) and a_kind_of(Z,f_18_aircraft)". The meaning list can be matched to parsed and interpreted user queries, so that only data matching the *meaning* of the user's query and not its exact words or syntax will be retrieved, saving considerably on the large times necessary to retrieve media data from secondary or tertiary storage.

This paper examines what can be learned indirectly about the things depicted in a pictorial multimedia datum based on its caption. We assume captions describe the important contents of the datum rather than evoking a theme or metaphor concerning the datum. We represent captions by meaning lists in storage. We have implemented a computerized expert system that will give one of six answers for a concept in a query meaning list when it is matched to a caption meaning list:

1. The query concept matches something in the datum.
2. The query concept definitely cannot match anything in the datum.
3. A part of the query concept matches something in the datum.
4. The query concept matches something in a set of data items.

5. A part of the query concept matches something in a set of data items.

6. The query concept may match something in the datum, but the caption is insufficient to tell, and the user must examine the datum.

Type inferences from natural-language descriptions are the best-known kind of inference (as in Allen (1987) and Grosz et al (1987) for artificial-intelligence applications, and Driscoll et al (1991), Rau (1987), Sembok and Rijsbergen (1990), and Smith et al (1989) for information-retrieval applications). For instance, a query requesting an airplane should match the caption "F-18 fighter" because F-18s are a type of airplane; a directed acyclic graph (DAG) of types can specify such relationships. We will investigate here other inferences. For instance, consider this actual caption from the database we have been studying of the Photo Lab of the Naval Air Warfare Center (NAWC), China Lake, California, USA:

F-15 USAF #74120, WA on tail, from Nellis AFB on apron with hangars in background and rain on runway.

Intuitively we can infer that the picture shows all of a F-15 aircraft with a "WA" visible on its tail fin; at least part of an apron, some hangars, and a runway; but probably not Nellis Air Force Base nor rain, and certainly not an explosion. Furthermore, we could look at the picture to determine whether the flaps are down or whether Hangar 3 is visible; looking would not determine the date or fuel load of this aircraft.

We will formalize here some of these inferences, especially database-independent inferences, in a manner analogous to that of the abovementioned papers. DiManzo et al (1986), Nagel (1988), and Wazinski (1992) have discussed some of these issues, but their approaches have not explicitly addressed useful implementations. Srihari and Rapaport (1989) does describe an implementation for the special case of identifying faces in photographs using captions.

2. Using caption syntax

2.1. Subject noun phrases

Since descriptive captions summarize the content of the corresponding media datum, the primary subject noun phrase usually denotes the most significant information in the media datum or its "focus". Thus we should expect that a portion of the space depicted by the media datum corresponds to the primary noun phrase. For instance in "Sidearm missile firing from AH-1J helicopter," we should expect to see a Sidearm missile firing, probably manifesting as flames and smoke, and probably in the center of the picture. On the other hand, we will not necessarily see a helicopter because it is in a prepositional phrase and is secondary in focus; this effect is more obvious in "Planes from the Enterprise" for which the planes may be far from the Enterprise.

The denotee of a subject noun phrase in a caption is not necessarily a large portion of the media datum in either space or time. In the preceding example, the flames and smoke may be small compared to the aircraft; however, they are unusual sights and thus in significance out of proportion to their size. Similarly, a caption for an audio tape of a birdcall may refer to only small portions of the tape in which the bird is actually singing. Thus rules are necessary to find it during content analysis.

A caption can have more than one subject noun phrase:

225 KTS seat ejection at 19 N 21 W. Rear dummy ejecting, drogue chute above but not open.

Here we have two sentences in which the second is effectively a compound sentence; the subjects are "ejection," "ejecting," and "chute", so all must be depicted in the picture. Multiple subjects can also be specified with "and" or other conjunctions, or often in the NAWC captions, with a "with" prepositional phrase:

Missile away from aircraft with plume and aircraft in view.

Here "missile," "plume," and "aircraft" are all depicted.

Appositives of the subject can serve to clarify an ambiguous subject by providing a better depictable name for it:

Missile, Sidewinder AIM-96, on stand.

Some appositives may require more work, but their referent can still can be identified:

Aircraft (tail 162966 and nose 87 with night hawk logo).

Gerunds can also serve as subjects, in which case they represent something more abstract but still depictable:

Shipboard weapons handling

Exterior building painting

Here we would expect to see some "handling" activity such as grasping or transporting for the first caption and paint equipment in the second.

Plural subjects imply that at least two objects of the indicated type must be present in the media datum. The quantification can be more precise, as in:

Two Sidewinder AIM-9M's

All three hangars in view

Both seat rockets with fire plume

2.2. A priori depictability

Being the subject of a sentence does not guarantee depiction, as in:

Cookoff test of Sidewinder AIM-9R.

View of Sidewinder mounted on F-18.

G-Range from the south.

Tests have no particular identifying physical manifestations; a view is another name for the media datum itself; and a large geographical area like G-Range has no identifying appearance. Other nondepictables are "program", "impact" of a missile, and "welcome". A priori depictability can be specified in the type hierarchy for terms that occur in the pictures. It inherits downward, so if tests are marked as depictable, sled tests inherit that property.

Depictability can also be a function of time and space. For instance, Sidewinder missiles did not exist

before 1950, so a picture dated before then cannot depict one. Usually depictability has no ending date, since obsolete equipment can be preserved for historical reasons.

2.3. Prepositional phrases

The subjects of prepositional phrases are not the caption focus. But most prepositions of physical relationship ("on", "in", "beside", etc. plus "of" when used for part-whole relationships) require that at least part of the object of the prepositional phrase must be depicted to thereby depict the relationship. For instance:

Harpoon missile launch aboard U.S.S. High Point PCH-1.

LGB, Skipper, bomb on MK 7 loader.

Nose of Harpoon missile.

MAJ John Gallinetti at the controls of the F/A-18 simulator at WSSA, Hangar 3.

In the first caption, we must see some of the ship; in the second, some of the loader; in the third, some of the missile; and in the fourth, some of the controls and some of the simulator. However, Hangar 3 at WSSA is made abstract by "at" instead of "inside".

There are multiword prepositions, like "off of." Some multiword phrases like "front of" and "top of" function as explicit acknowledgement that the named part of the prepositional object is depicted.

An important exception to partial depiction of prepositional objects is the "view of" phrase and its variants ("side view of", "picture of", "portrait of", "detail of", "view during", "graphics composite", etc.) which describe the datum as a whole. The object of their prepositional phrase is the true subject, and is necessarily depicted in full:

Excellent picture of the drone in the midst of explosion.

Closeup view of drone and aircraft wing.

"View from", however, takes as object an abstract location, and neither the view nor its object is depicted, as in:

Sidewinder AIM-9R on an F-18. View from underside of aircraft.

where the missile is the subject.

Some prepositional phrases can be depictable in themselves, like for "F-18 in the air" or "F-18 in takeoff", where the "in" phrase could be confirmed by examining the picture. These examples are different from "F-18 in review" and "F-18 in trouble". So preposition-object pairs need additional information in the type hierarchy to indicate for which prepositions they or their subtypes are depictable.

2.4. Adjectives

Adjectives often establish the subtype of their noun phrase, simplifying the finding of the referent. For example, missiles are many sizes and shapes, but "cruise missiles" are considerably more restricted, and "Tomahawk cruise missiles" have known exact dimensions. Thus some adjectives identify more specific referents in the type hierarchy.

Other adjectives associate specific physical things with abstract concepts, as in "Sidewinder separation test" and "steel railguide". Natural-language processing must distinguish these from known subtypes, so in "steel warhead cover" the warhead cover will be recognized as a cover subtype, but steel will be recognized as a material. A more complex example:

TP89053 air-to-air Sidewinder separation test.

Here "separation test", "air-to-air Sidewinder", and "TP89053 Sidewinder" are known subtypes, so the parsing should be ((TP89053 (air-to-air Sidewinder)) (separation test)).

Still other adjectives restrict properties of a noun, as "open cockpit" and "cold engine". Some like "open" can be confirmed by examining a picture even if not mentioned in the caption, like some prepositional phrases.

2.5. Verbs and verbals

Verbs and verbals pose problems for depiction analysis since they tend to be more abstract than nouns

and adjectives. A still photograph can only depict the helicopter in a single position; video is the only way to directly depict "taking off". But a multimedia database cannot always employ the best media.

Participles are the most common verbal in the NAWC captions. Present-tense participles tend to be depictable, while past participles are not because they suggest actions already done. For instance:

YA-4C BU#145063 aircraft firing HIPEG MK-4 gun pod.

Parachutes opened on four weapons.

The firing is visible, but the opening is over.

Adverbs usually specify subtypes of the verb's action, analogous to adjectives restricting nouns:

Missile on aircraft, just igniting.

Air-to-air Skipper AGM-123A double firing from A-6E BU#155592 aircraft.

But just as with adjectives, some adverbs assert non-subtype properties of the verb which may be depictable by themselves, like in "missile trailing behind" and "missile mounted parallel".

Direct objects of verbs are usually depicted because they are usually essential to understanding the verb. For instance:

Sidewinder hitting drone.

Personnel assembling ABL motors.

where "hitting" and "assembling" would not be interpretable without the drone and motors. But in the YA-4C caption above, the gun pods are not necessarily visible since the firing can be visible alone. Some adverbial prepositional phrases are similar:

Helicopter taking off from pad.

where the pad must be visible in part to make "taking off" clear in the picture. This depictable adverbial "from" object is different from the nondepictable adjectival "from".

2.6. Depiction-cancelling modifiers

Modifiers of a noun or verbal can override depiction:

F-4 on fire

F-4 during assembly

In the first, fire means unusual major changes in appearance, so that the object may no longer be visible. In the second, "assembly" contradicts the usual shape implications of an F-4. Fortunately, there are not many of these exceptions, and they can be enumerated.

3. Additional inferences from captions

3.1. Caption vagueness

Too-general words in a caption description imply answer 6 of section 1, "look and see". For instance, query term "Sidewinder" might match caption "missiles on an F-18" after inspection of the picture. But our parser does eliminate redundant overgeneral caption terms, like the first and third words of "Sidewinder AIM-9R missile".

3.2. Part-whole inferences

Depiction of a part implies the partial depiction of the whole, answer 3 of section 1. For instance, a caption that mentions as subject an "F-18 wing pylon" implies that all of the pylon can be seen and also that part of an F-18 wing and part of an F-18 can be seen. There are some exceptions related to removable parts: an automobile tire is not necessarily on an automobile, but the front of a missile must be on a missile.

On the other hand and contrary to intuition, depiction of a whole does not always imply depiction of the part. For instance, a "view of a missile" caption implies depiction of that missile, but does not necessarily the front of the missile, its left fin, etc. There are exceptions, however, for (1) parts visible in every view, even a partial view, like the fuselage of an airplane or the finish on a desk, and (2) parts necessary to its identity, like the head of a person. Exceptions can be marked in the type hierarchy.

Part-whole relationships can be generalized; for instance, a fin is part of a Sidewinder because a Sidewinder is a kind of missile, a tail fin is part of any missile, and a tail fin is a kind of fin. Such inferences can be formalized in the following axioms, exploited in our implementation. Here "a_kind_of(X,Y)" means X is a kind of Y, "part_of(X,Y)" means X is part of Y, and "depicted(X,T)" means X is depicted in the picture with type T of section 1.

3.3. Category-compatibility inferences

Media data usually can be categorized by intended use. In the NAWC database, pictures categories are:

P1: views of a whole product (like a missile)

P2: assembly and loading

P3: tests (the largest category)

P4: vehicles

P5: facilities

P6: events involving people (including historical)

P7: portraits of people

P8: non-mission (including geography, flora, and fauna)

Keywords help identify the category for a caption. For instance, any mention of sleds puts the caption in P3; mention of a stand puts it in P1; and mention of people puts it in P6 or P7. But in general, category classification requires parsing and interpretation of the caption because patterns of concepts, not any single concept, best establish the category, like the co-occurrence of "missile" and "firing"

establishing category P3. Thus classification requires domain-dependent rules written by a domain expert, but they are generally easy to create for a multipurpose database because they often correspond to simple intuitions well understood by users, as above.

Queries can also be assigned a corresponding picture type or types. Then if a query and caption have disjoint sets of types, then the depictables of the query cannot match the caption (answer 2 of section 1), and no further work is required.

3.4. Importance inferences

A database has purposes, and certain things associated with those purposes must be mentioned when present in the datum. For instance, the NAWC photographs usually show weapons, so a weapon not mentioned can be inferred absent (answer 2 of section 1). So a query "F-14 with flaps down" should not match the caption "F-18 with flaps down", but could match "F-14 in flight". Like picture category inferences, this requires domain-dependent identification of high-importance concepts by a domain expert, and flagging of them in the type hierarchy, but this should be easy (e.g. NAWC's official name mentions weapons). Furthermore, the inferences can be made less arguable by making them conditional on category.

3.5. Ellipsis inferences

If order of the data is known, as with the NAWC numbering scheme, then captions may be put into correspondence to suggest missing repeated information (ellipsis) in the last. For instance, if three successive captions mention tests of Sidewinder missiles, and the next caption mentions a test but no subject, then assume it is a Sidewinder.

4. Supercaptions

36% of the captions in the NAWC database are "supercaptions", captions describing more than one

media datum. They help minimize caption redundancy. For instance:

BLA, vertical launch ASROC jet vane actuator. Various views.

Supercaption information can also be implicit within single-datum captions, in leading repeated phrases:

Exterior building painting. Bldg. 00499 BOQD Lexington.

Exterior building painting. Bldg. 30954 Assembly Building G-2 range.

Such leading phrases occur in about 18% of the NAWC captions and supercaptions.

Such simple supercaption information can be considered to be implicitly universally quantified over the set of all subcaptions. That is, it is appendable to the subcaptions. So all subcaptions of the BLA caption must show a BLA; "various views" is not depictable.

Supercaptions can also include differential semantics for the subcaptions, or explanations of how the sub-datums differ, and these are not universally quantified. "Various views" is an example, since the set of pictures shows various views, not each picture individually. We can infer this from the principle that photographs usually show only one view of something. Contradictory supercaption information often signals differential semantics:

TCP-E Harm no. 1 motor fast cookoff. Pre and post test views.

Since a view cannot be both pre-test and post-test, some pictures must be one and some must be the other. Answer 4 of section 1 is appropriate for a query requiring "pre-test". In the NAWC captions, differential information is usually confined to a single sentence per caption to avoid confusion.

Differential semantics can suggest complex correspondences. Consider:

S040, Agile, WAGOS Seeker natural vibration test. Overall, closeup of front, accelerometers in turn position, and tail view in room 123.

The second sentence contains contradictory view information, so we suspect that it is differential information. If this supercaption describes four pictures, the four phrases in it must correspond to the pictures. In addition, if the four picture numbers are consecutive that suggests that the four phrases correspond the pictures in order.

Ambiguity in correspondences can occur in differential information, as in "Sled on track and various views of target" describing six pictures. Here at least one picture must be of the sled on track, probably the first, but we know little about the other pictures. In cases like this it may be tedious for the caption writer to specify all the details. Thus, data retrieval for a query matching such supercaptions may include datum sets in which at least one answer to the query will be found, but we are not sure which datum it is. These are answer types 4 and 5 of section 1.

5. A larger example

To show how these ideas are put together, consider the following caption:

> F-15 USAF #74120, WA on tail, from Nellis AFB on apron with hangars in background and rain on runway. Full side, rear, and front views. Excellent series.

This refers to a set of four photographs. A particular F-15 aircraft is fully depicted since it is the subject of the first sentence. A code marking of "WA" is also fully depicted since it is an appositive, and the tail of the aircraft on which the code resides is at least partially depicted because it is related via an "on" relationship. Nellis AFB is not depicted because it is related via the wrong sort of prepositional connection, but the apron is depicted because it uses another "on" relationship. The hangars and rain are also depicted at least in part because they are connected while the special co-agent preposition "with," and the runway is also visible, but the "background" is not because it is not depictable regardless of the "in." "Hangars" means at least two must be depicted.

As for the second sentence, there are three contradictory views specified, so this must be information that must be assigned to subsets of the pictures. Since there are four pictures, and only three of these contradictory descriptors, we do know which or how many of the pictures have the descriptors. But following the heuristic that such contradictory descriptions usually follow the order of the pictures, we can assume that the first picture is a full side view and the last picture a front view; the second is either a full side or a rear, and the third is either a rear or a front. Once this assignment of the contradictory information has been sorted out, we can make other inferences; for instance, the side of the F-15 must

be depicted in the first picture. The third sentence refers to the photographic quality of the entire set of pictures and since "series" is undepictable, the third sentence does not add any other depiction information.

As for indirect inferences of depiction, we can infer that there is no second aircraft in any of the pictures. Furthermore, we can infer that there are no weapons testing activities in any of the pictures. We should expect to see part of the apron and part of the runways, and not necessarily all of them. We may be able to confirm an asphalt runway by examining the picture. Testing and people are not mentioned, and an airplane is mentioned, so the picture type is P4 (vehicles).

6. Integrating content analysis

The vagueness of the type-6 query answers of section 1 can be reduced by computerized content analysis of the data. This means building an analogous meaning list for the datum with visual processing techniques for pictures, signal processing techniques for signals, etc. Then the caption meaning list can be enriched with the contents of the content-analysis meaning list. This is future work for us.

7. Experiments

We implemented our ideas of sections 2-4 as a rule-based expert system in Quintus Prolog. We took 23 test queries (21 of which were supplied by the NAWC Photo Lab as typical) and the random sample of 204 captions and 2 supercaptions studied in Rowe and Guglielmo (1992), captions containing 830 distinct nouns and verbs. The two queries we wrote were vaguer and more complicated in syntax than the others, and provided a different challenge. We parsed and interpreted the queries and captions using our current software, and ran our expert system on the resulting meaning lists and the parser's type hierarchy. Each query was compared to each caption.

Our expert system inferred 868 type-1 term matches, 3411 type-2, 368 type-3, 5 type-4, 0 type-5, and 810 type-6. To estimate term-match precision, we randomly selected from the output 230 query-caption

pairs containing 274 matches of all types except 5, and examined the corresponding picture. This grading was easy, as the terms were easily recognizable pictorially. Of the 274 sampled term matches, 14 were not justifiable, giving a precision of 0.95. Of the 14, 10 resulted from inadequate natural-language processing, and 4 from inadequate specialized knowledge (like that aircraft appear at an airport). Since there were only 41 exact matches between query terms and caption terms, term match recall increased by a factor of 48 counting only positive answers, and 126 counting all answers.

Precision and recall for the records retrieved require some caveats. Since we are providing answers to new kinds of questions with all but type-1 matches, it may be misleading to compare our system to traditional information retrieval; our ideas will better support usage like hypertext browsing, in suggesting possibly-related data to explore. Match metrics and thresholds will need to be different for this new kind of usage (for instance, we must judge the value to the user of seeing all of an aircraft instead of its wing).

Nonetheless, we did estimate overall recall and precision for our system (see Table 1). We manually examined the 23*204 = 4692 possible *query-picture pairs excluding supercaptions*. Of the 4692, we decided that 76 pairs were perfect matches, using the captions to help understand what the pictures showed. We then extracted the nouns and verbs of the queries, compiled keyword lists (averaging 3.7 words per query), and matched the 23 lists to the 204 captions, removing the standard noun and verb suffixes; experiments 1-4 show the resulting precision and recall for different threshold percentages for keyword matching of the keywords in the query. We then ran the software described in this paper on the noun and verbs concepts automatically extracted from every query and caption meaning list, and compared all query-caption pairs in experiments 5-11, excluding those with at least one type-2 (negative) match. Experiments 5-7 required exact matches for every noun and verb concept in the query, whereas experiments 8-10 accepted a 65% or more match percentage. Experiments 5 and 8 considered only query-caption pairs in which term matches are type-1 (logically certain) as described in section 1; experiments 6 and 9 considered matches of either type-1 or type-3 (match of a part); and experiments 7 and 10 considered matches of either type-1, type-3, or type-6 ("look and see" matches). Finally, experi-

ment 11 examined all pairs without a type-2 match.

Our expert system required only 137 additional rules (Prolog subroutines) beyond the natural-language lexicon, parsing, and interpreting. Experiments 1-3 took an average of 0.66 seconds of CPU time on a Sun Sparcstation per query-caption pair for an implementation without indexing; experiments 5-10 took 0.34 seconds per pair; and experiments 4 and 11 required negligible processing. Thus our ideas need little time. In addition, experiments 5-10 required precomputation of meaning lists, needing about 3 seconds per query and 10 seconds per caption (the latter of which can be done long before querying). But experiments 1-3 required keyword lists from the user, which require more thinking time than the English phrases needed for experiments 5-10.

Our results show significant improvements in recall for a fixed precision over keyword matching when the 65% match criterion is used for query terms. Compare experiments 3 and 10 in particular: we got 60% more hits in retrieving 37% fewer data items. And the four items missed in experiment 10 had poorly written captions. Comparing experiments 4 and 11 also shows that our negative inferences can significantly help when a user has trouble formulating a good query. But our results are less impressive for a 100% match criterion. Note that our experiments did not do full matching (including structural matching) of query meaning list to caption meaning list, the "fine-grain" match done in Rowe and Guglielmo (1993), and further improvements in recall and precision are likely to accrue from using it.

8. Conclusion

Multimedia data can require substantial processing time. We have shown some important ways in which query-based access to a multimedia database can be intelligently analyzed to reduce unnecessary data accesses and suggest matches not literally confirmed. Our methods exploit simple properties of the query and a simple set of inference rules. Our methods should be helpful for the many multimedia databases in which captions describe (rather than discuss) the contents of the media data.

9. References

Allen, J. (1987). *Natural Language Understanding*. Menlo Park, CA: Benjamin Cummings.

DiManzo, M., Adorni, G., and Giunchiglia, F. (1986, July). Reasoning about scene descriptions. *Proceedings of the IEEE, 74(7)*, 1013-1025.

Driscoll, J., Ragala, D., Shaffer, W., & Thomas, D. (1991). The operation and performance of an artificially-intelligent keywording system. *Information Processing and Management, 27(1)*, 43-54.

Grosz, B., Appelt, D., Martin, P., & Pereira, F. (1987). TEAM: An experiment in the design of transportable natural language interfaces. *Artificial Intelligence, 32*, 173-243.

Nagel, H. (1988, May). From image sequences towards conceptual descriptions. *Image and Vision Computing, 6(2)*, 59-74.

Rau, L. (1987). Knowledge organization and access in a conceptual information system. *Information Processing and Management, 23(4)*, 269-284.

Rowe, N. & Guglielmo, E. (1993). Exploiting captions in retrieval of multimedia data. *Information Processing and Management*, to appear.

Sembok, T. & van Rijsbergen, C. (1990). SILOL: A simple logical-linguistic retrieval system. *Information Processing and Management, 26(1)*, 111-134.

Smith, P., Shute, S., Galdes, D., & Chignell, M. (1989, July). Knowledge-based search tactics for an intelligent intermediary system. *ACM Transactions on Information Systems, 7(3)*, 246-270.

Srihari, R. & Rapaport, W. (1989, November). Combining linguistic and pictorial information: using captions to interpret newspaper photographs. In *Lecture Notes in Artificial Intelligence 437: Current Trends in SNePS--Semantic Network Processing System*, Kumar, D., ed., Proceedings of First Annual SNePS Workshop, Buffalo NY, 85-96.

Wazinski, P. (1992, April). Generating spatial descriptions for cross-modal references. Third Conference on Applied Natural Language Processing, Association for Computational Linguistics, Trento, Italy, 56-63.

Experiment Number	*Threshold*	*Notes*	*Number of Items Fetched*	*Number of Hits*	*Precision*	*Recall*
1	100% of keywords		22	20	0.91	0.26
2	>65% of keywords		42	25	0.60	0.33
3	>49% of keywords	And >1 keyword	295	45	0.15	0.59
4	none		4692	76	0.016	1.00
5	100% of meaning list terms	Type-1 matches only	22	16	0.73	0.21
6	100% of meaning list terms	Type-1 and Type-3 matches only	33	27	0.82	0.36
7	100% of meaning list terms	Type-1, Type-3, and Type-6 matches only	47	28	0.60	0.37
8	>65% of meaning list terms	Type-1 matches only	31	19	0.61	0.25
9	>65% of meaning list terms	Type-1 and Type-3 matches only	58	40	0.69	0.53
10	>65% of meaning list terms	Type-1, Type-3, and Type-6 matches only	186	72	0.39	0.95
11	none	Excluding Type-2 matches	2353	76	0.032	1.00

Table 1: Experimental results for 4692 query-caption pairs of which 76 were manually judged as perfect matches ("hits")

Distribution List

Defense Technical Information Center Cameron Station Alexandria, VA 22314	2
Library, Code 52 Naval Postgraduate School Monterey, CA 93943	2
Center for Naval Analyses 2000 N. Beauregard Street Alexandria, VA 22311	1
Director of Research Administration Code 08 Naval Postgraduate School Monterey, CA 93943	1
Mr. Russell Davis HQ, USACDEC Attention: ATEC-1M Fort Ord, CA 93941	2
Dr. Neil C. Rowe, Code CSRp Naval Postgraduate School Computer Science Department Monterey, CA 93943	50
Prof. Ted Lewis, CS/Lt Naval Postgraduate School Computer Science Department Monterey, CA 93943	2